AIT 614 PROJECT PROPOSAL

**ANALYZING SUICIDE RATES: A DATA-DRIVEN APPROACH**

TEAM 8: NEELIMA PALLEBOINA, KRISHNA REDDY

PPROFESSOR’S NAME: Lindi Liao

GEORGE MASON UNIVERSITY

COURSE NAME: AIT 614 SECTION 001

DATE: 11/06/2023

Analysing suicide rates is a critical task, as it can provide insights into the factors contributing to suicide and inform prevention efforts. Here's a general outline for an analysis of suicide rate data.

**INTRODUCTION**:

Importance of analyzing suicide rates and patterns and their importance in public health: Analyzing suicide rates and patterns is of paramount importance in public health. By dissecting the data on suicide, public health officials gain crucial insights into a deeply complex and devastating issue. This analysis allows for the identification of high-risk populations, such as specific age groups, genders, and socio-economic strata, enabling the implementation of targeted intervention strategies. Furthermore, the ability to detect trends and changes in suicide rates over time is instrumental in addressing emerging problems and assessing the effectiveness of prevention programs.

We analysed the suicide rates of all the countries from 1985 to 2016. Our goal is to perform a broad examination of suicide rates in all the countries by sex, race, Hispanic origin, and age. We will provide a detailed explanation of these statistics in our project.

**DATA COLLECTION:**

We checked few open sources to find the dataset related to suicide rates, few of the open sources which we used are Centres for Disease control and prevention[1] , Data.gov [2], U.S Department of Health & Human services. We also analysed the data from common sources which include the World Health Organization (WHO), national health agencies, and research institutions to know why there is a rise in the suicide rates in developing and also in developed countries.

**VARIABLES IN THE DATASET:**

Country: Countries all over the world arranged in alphabetical order

Year: year from 1985 to 2016

Sex: categorized into male and female

Age: Categorized into 5 age groups: 15-24, 25-34, 35-54, 55-74, 75+

Suicide\_no: Number of suicides that took place in a particular year and country

Population: overall population of the country

Suicide/100k pop: suicides per 100k of the population in a specific country

Country-year: country and the year that a suicide has took place

gdp\_per\_year: GDP of a country taken per year

gdp\_per\_capita: GDP of country taken per capita

generation: Age group of the people which is categorized into 5 : Generation X, Silent, G.I Generation, Boomers, Millennials.

**DATASET:**

Suicide dataset has 11 columns and 27821 rows which has suicide rates of each country arranged in alphabetical order.

**A table with numbers and numbers

Description automatically generated**

**DATA CLEANING AND PRE-PROCESSING:**

Cleaning the data by addressing missing values, outliers, and inconsistencies is yet to be done. There are few missing values or null values for few of the countries which we have to clean the dataset. There is also no data available for the suicide rates that took place for few years in some countries.

We will normalize or standardize variables if necessary.

**RESEARCH WORK:**

We found two related research works on the similar dataset and topic.

**RESEARCH WORK 1:**

In one of the research work they started with Exploratory data analysis using matplotlib, They performed the distribution graphs by taking two columns data for each histogram and bar graph and for visualization purpose they took the columns that has the unique values between 1 and 50.

They dropped the columns with Nan values which we will be performing too in our project. And they also checked if any column has more than one unique value, which we felt was not necessary.

They performed correlation matrix, scattered and density plots for all the columns to find the patterns and trends in the data.

**RESEARCH WORK 2:**

In the other research work they used the libraries pandas, seaborn, matplotlib, numpy, scipy, pycountry, and geopandas to do Exploratory data analysis. The dataset is loaded into a pandas DataFrame, and the column names are renamed for clarity.

They wrote a code to analyzes suicide rates based on sex, age, and generation.

Plots are generated to visualize trends in suicide rates among men and women, different age groups, and various generations.

Next they wrote a code which prepares the data for mapping suicide rates by country.

They used geopandas to merge the dataset with geographical information and plots the suicide rates on a world map.

The code explores the relationship between suicide rates and GDP per capita for different countries.

They calculated the correlation coefficient and generated regression plots.

Then they did a further analysis on countries with high and low correlation coefficients.

The overall analysis in this research work involved exploring the correlation between suicide rates and various factors, such as age, generation, and GDP per capita, and visualizing the results using various plots and maps.

**RESEARCH WORK 3:**

In one of the research works they plotted distribution Graphs Histogram and Bar Graphs

The function plotPerColumnDistribution is used to generate distribution graphs for columns with between 1 and 50 unique values. For numerical columns, it plots histograms; for categorical columns, it plots bar graphs.

The graphs which they plotted provided insights into the distribution of values in each selected column.

Correlation Matrix:

The function plotCorrelationMatrix generates a correlation matrix for the dataset.

They dropped the columns with NaN values and kept the columns with more than one unique value.

For the resulting matrix is visualized using a heatmap, showing the correlation coefficients between pairs of columns. This helps identify potential relationships between variables.

Scatter and Density Plots:

The function plotScatterMatrix creates scatter plots for numerical columns and kernel density plots along the diagonal.

It provides a visual overview of the relationships between pairs of numerical variables, with correlation coefficients annotated on the plots.

Overall, these visualizations are essential for understanding the distribution of individual variables, detecting patterns or correlations, and gaining insights into the overall structure of the dataset. They serve as a starting point for more in-depth analysis and model building.

**OBJECTIVES OF THE PROJECT:**

* Our primary aim is to analyse the suicide patterns and causes.
* Gain insights into primary areas leading to higher suicide rates.
* To analyse if GDP of the country is affecting the suicide rates.
* Investigate regional and country-specific variations in suicide rates.

In our project we’ll perform a more comprehensive analysis of suicide rates compared to research work 1 and 2 using following steps:

**Feature Engineering:**

1. Time-Based Features:

We are planning to extract additional time-based features, such as month and season, to explore whether there are specific patterns or trends during certain times of the year.

2. Combine Gender and Age:

We’ll Create a combined feature that represents both gender and age group to capture interactions between these variables.

**Missing Data Handling:**

1. Handle Missing Values:

Examine and handle missing values in the dataset, especially in the "HDI for year" column. We are planning to drop columns with a high proportion of missing data.

**Exploratory Data Analysis (EDA):**

We want to perform the exploratory data analysis to:

* Analyse suicide rates by gender and age group.
* Investigate regional and country-specific variations in suicide rates.
* Identify trends and patterns in the data.

1. Correlation Analysis:

We will Conduct correlation analysis to identify relationships between different variables, including the correlation between suicide rates and other socio-economic factors like GDP per capita.

2. Data Distribution:

Explore the distribution of suicide rates across different countries, genders, and age groups. [5]

**Machine Learning Models:**

1. Regression Models:

We will use two regression models linear regression, random forest regression to predict the suicide rate based on various features. This will help us to identify significant predictors and what variables are affecting the suicide rates the most.

2. Classification Models:

We are planning to convert suicide rates into classes (low, medium, high) and apply two classification models logistic regression, decision trees to classify countries into these classes based on their features and plot the graphs accordingly.

**Time Series Analysis:**

1. Time Series Forecasting:

As the dataset spans in multiple years, we are considering to use time series forecasting models to predict future suicide rates.

**Clustering:**

1. Country Clustering:

We will apply clustering algorithms to group countries based on similar suicide rate trends. This will help us to reveal patterns among countries with similar socio-economic characteristics.

**Model Interpretability:**

1. Feature Importance:

We will try to include the models which can provide us feature importance scores to understand which variables contribute the most to predicting suicide rates.

**Visualization:**

1. Interactive Visualizations:

We will plot interactive visualization to present the analysis in a more user-friendly and accessible manner.

**Hypothesis Testing:**

1. Statistical Tests:

We will Perform statistical tests to validate hypotheses about the relationships between variables, such as the impact of GDP per capita on suicide rates.

**DIAGRAM OR DESCRIPTION OF PROPOSED SYSTEM:**

A diagram of a project

Description automatically generated

**PROJECT TASKS AND TIMELINE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **TASKS** | **ASSIGNED TEAM MEMBER** | | **TIMELINE** |
| Analysing the research work | Neelima | | 11/05/2023 - 11/06/2023 |
| Data cleaning and Data pre-processing | Neelima | | 11/07/2023 – 11/06/2023 |
| Exploratory Data Analysis | Krishna | | 11/08/2023 – 11/16/ 2023 |
| Graphical Analysis | Neelima | | 11/17/2023 – 11/19/2023 |
| Correlation Analysis | Krishna | | 11/20/2023 – 11/23/2023 |
| Time series Analysis | Krishna | 11/24/2023 – 11/26/2023 | |

**EXAMPLE GRAPH:**

Here we analyzed the number of suicides took place in each country all the years together.

A graph with different colored bars

Description automatically generated

**CONCLUSION:**

The insights and knowledge gained from the analysis can be integrated with public policy systems and medical institutions for further research. New behavioral science methods can be implemented in psychology fields.

The data we get in the form of graphs, prediction accuracy, and patterns can be used by schools, companies(since most rates are recorded by age 24-44), and other awareness platforms.

**REFERENCES:**

[1]<https://www.cdc.gov/suicide/suicide-data-statistics.html>

[2] <https://catalog.data.gov/dataset?tags=suicide>

[3]<https://www.researchgate.net/publication/20917681_Time-series_analyses_of_the_American_suicide_rate>

[4] <https://www.cdc.gov/surveillance/blogs-stories/Suicide-Trends.html>

<https://data.oecd.org/healthstat/suicide-rates.htm>

<https://www.kaggle.com/code/dornani/a-classification-analysis-on-suicide-data>

<https://medium.com/analytics-vidhya/exploratory-data-analysis-eda-in-python-on-suicide-rates-33da65855add>

[5] <https://towardsdatascience.com/the-suicide-crisis-in-data-7025f8551ca8>

<https://nycdatascience.com/blog/student-works/data-analysis-on-suicide-rates-across-the-globe/>

<https://datauab.github.io/suicide-rates/>